Mutation Annotation Format (MAF) File Specification

Mutation annotation files should be transferred to the DCC. Those files should be formatted using the mutation annotation format (MAF) that is described below. The file names should have the suffix "maf" and contain the prefix of the containing archive name (*e.g.* genome.wustl.edu_OV.IlluminaGA_DNASeq.1.maf). The serial number in the name (*e.g.* the 1 in the previous file name) is no longer tied to the archive (*i.e.* it can be any integer) so that multiple MAF files can exist in the same archive. You can also add optional metadata in the file name between the platform and serial number (e.g. genome.wustl.edu_OV.IlluminaGA_DNASeq.prelimiary.1.maf)

The following data are reported in MAF files:
Somatic mutations

- Missense and nonsense
- Splice site, defined as SNP within 2 bp of the splice junction
- Silent mutations
- Indels that overlap the coding region or splice site of a gene or the targeted region of a genetic element of interest.
- Frameshift mutations
- Mutations in regulatory regions

SNPs

- Any germline SNP with validation status "unknown" is included.
- SNPs already validated in dbSNP are not included since they are unlikely to be involved in cancer.

Validation
All candidate somatic missense, nonsense, splice site and indels are retested by an independent (orthogonal) genotyping method. If the SNP is confirmed by an independent method, they are deemed valid. Silent mutations may be validated for the purpose of calculating the background mutation rate. No germline (SNP or indel) candidates are processed through validation. However, if the validation process reveals a given candidate somatic variation event to be germline or loss of heterozygosity, those validated data are reported in the validation file.

A *validated somatic mutation* is identified by (Verification_Status=Verified or Validation_Status=Valid) and Mutation_Status=Somatic.

MAF files have a base data type of "Mutations". Putative (un-validated) somatic mutations or non-somatic mutations are considered Level 2 data and are available as controlled access only. Validated somatic mutations (defined above) are considered Level 3 data and open access.

Mutation Annotation Format File Fields

The format of a MAF file is tab-delimited columns. Those columns are described in Table 1 and are required in every MAF file. The order of the columns will be validated by the DCC. Column headers and values **are** case sensitive where specified. Columns may allow null values (*i.e.* blank cells) and/or have enumerated values. The validator looks for a header stating the version of the specification to validate against (e.g. #version 2.0). If not header is present the validator assumes the MAF file is version 1.x. Any columns that come after the columns described in Table 1 are optional. Optional columns are not validated by the DCC and can be in any order.

Table 1 - Mutation annotation format (MAF) version 2.0 file column headers

Index	MAF Column Header	Description of Values	Example	Case Sensitive	Null	Enumerated
1	Hugo_Symbol	HUGO symbol for the gene (HUGO symbols are *always* in all caps). If no gene exists within 5kb enter "Unknown". Source: http://genenames.org	EGFR	Yes	No	Set or Unknown
2	Entrez_Gene_Id	Entrez gene ID. Source: http://ncbi.nlm.nih.gov/sites/entrez?db=gene	1956	No	No	Set
3	Center	Genome sequencing center reporting the variant. If multiple institutions report the same mutation separate list using semicolons.	hgsc.bcm.edu	Yes	No	hgsc.bcm.edu, broad.mit.edu, or genome.wustl.edu
4	NCBI_Build	NCBI human genome build number with decimal.	36.1, 37.0, etc.	No	No	Set
5	Chromosome	Chromosome number without “chr” prefix that contains the gene.	X, Y, M, 1, 2, etc.	Yes	No	Set
6	Start_Position	Lowest numeric position of the reported variant on the genomic reference sequence. Mutation start coordinate (1-based coordinate system).	999	No	No	Set
7	End_Position	Highest numeric genomic position of the reported variant on the genomic reference sequence. Mutation end coordinate (inclusive, 1-based coordinate system).	1000	No	No	Set
8	Strand	Genomic strand of the reported allele. Variants should always be reported on the positive (+) genomic strand.	+	No	No	+ or -

9	Variant_Classification	Translational effect of variant allele.	Missense_Mutation	Yes	No	Frame_Shift_Del, Frame_Shift_Ins, In_Frame_Del, In_Frame_Ins, Missense_Mutation, Nonsense_Mutation, Silent, Splice_Site_Del,Splice_Site_Ins, Splice_Site_SNP, Nonstop_Mutation, 3'UTR, 3'Flank, 5'UTR, 5'Flank, IGR, Intron, RNA, or Targeted_Region
10	Variant_Type	Type of mutation. TNP (tri-nucleotide polymorphism) is analogous to DNP but for 3 consecutive nucleotides. ONP (oligo-nucleotide polymorphism) is analogous to TNP but for consecutive runs of 4 or more.	INS	Yes	No	SNP, DNP, TNP, ONP, INS, DEL, or Consolidated
11	Reference_Allele	The plus strand reference allele at this position. Include the sequence deleted for a deletion, or "-" for an insertion.	A	Yes	No	A,C,G,T, and/or -
12	Tumor_Seq_Allele1	Primary data genotype. Tumor sequencing (discovery) allele 1. "-" for a deletion represent a variant. "-" for an insertion represents wild-type allele. Novel inserted sequence for insertion should not include flanking reference bases.	C	Yes	No	A,C,G,T, and/or -
13	Tumor_Seq_Allele2	Primary data genotype. Tumor sequencing (discovery) allele 2. "-" for a deletion represents a variant. "-" for an insertion represents wild-type allele. Novel inserted sequence for insertion should not include flanking reference bases.	G	Yes	No	No
14	dbSNP_RS	Latest dbSNP rs ID (dbSNP_ID) or "novel" if there is no dbSNP record. source: http://ncbi.nlm.nih.gov/projects/SNP/	rs12345	Yes	Yes	Set or "novel"

15	dbSNP_Val_Status	dbSNP validation status. Semicolon-separated list of validation statuses.	by2Hit2Allele;byCluster	No	Yes	by1000genomes;by2Hit2Allele; byCluster; byFrequency; byHapMap; byOtherPop; alternate_allele
16	Tumor_Sample_Barcode	BCR aliquot barcode for the tumor sample including the two additional fields indicating plate and well position. i.e. TCGA-SiteID-PatientID-SampleID-PortionID-PlateID-CenterID. The full TCGA Aliquot ID.	TCGA-02-0021-01A-01D-0002-04	Yes	No	Set
17	Matched_Norm_Sample_Barcode	BCR aliquot barcode for the matched normal sample including the two additional fields indicating plate and well position. i.e. TCGA-SiteID-PatientID-SampleID-PortionID-PlateID-CenterID. The full TCGA Aliquot ID; e.g. TCGA-02-0021-10A-01D-0002-04 (compare portion ID '10A' normal sample, to '01A' tumor sample).	TCGA-02-0021-10A-01D-0002-04	Yes	No	Set
18	Match_Norm_Seq_Allele1	Primary data. Matched normal sequencing allele 1. "-" for deletions; novel inserted sequence for INS not including flanking reference bases.	T	Yes	Yes	A,C,G,T, and/or -
19	Match_Norm_Seq_Allele2	Primary data. Matched normal sequencing allele 2. "-" for deletions; novel inserted sequence for INS not including flanking reference bases.	ACGT	Yes	Yes	A,C,G,T, and/or -
20	Tumor_Validation_Allele1	Secondary data from orthogonal technology. Tumor genotyping (validation) for allele 1. "-" for deletions; novel inserted sequence for INS not including flanking reference bases.	-	Yes	Yes	A,C,G,T, and/or -
21	Tumor_Validation_Allele2	Secondary data from orthogonal technology. Tumor genotyping (validation) for allele 2. "-" for deletions; novel inserted sequence for INS not including flanking reference bases.	A	Yes	Yes	A,C,G,T, and/or -

22	Match_Norm_Validation_Allele1	Secondary data from orthogonal technology. Matched normal genotyping (validation) for allele 1. "-" for deletions; novel inserted sequence for INS not including flanking reference bases.	C	Yes	Yes	A,C,G,T, and/or -
23	Match_Norm_Validation_Allele2	Secondary data from orthogonal technology. Matched normal genotyping (validation) for allele 2. "-" for deletions; novel inserted sequence for INS not including flanking reference bases.	G	Yes	Yes	A,C,G,T, and/or -
24	Verification_Status	Second pass results from independent attempt using same methods as primary data source. Generally reserved for 3730 Sanger Sequencing.	Verified	Yes	Yes	Verified, Unknown
25	Validation_Status	Second pass results from orthogonal technology.	Valid	Yes	Yes	Valid, Unknown, Wildtype
26	Mutation_Status	Updated to reflect validation or verification status.	Somatic	Yes	No	Somatic, Germline, None, LOH, or Unknown
27	Sequencing_Phase	TCGA sequencing phase. Phase should change under any circumstance that the targets under consideration change.	Phase_I	No	Yes	No
28	Sequence_Source	Molecular assay type used to produce the analytes used for sequencing.	PCR;Capture	Yes	No	PCR, Capture, WGS
29	Validation_Method	The assay platforms used for the validation call. Examples: Sanger_PCR_WGA, Sanger_PCR_gDNA, 454_PCR_WGA, 454_PCR_gDNA; separate multiple entries using semicolons.	Sanger_PCR_WGA;Sanger_PCR_gDNA	No	Yes	No
30	Score	Not in use.	NA	No	Yes	No
31	BAM_File	Not in use.	NA	No	Yes	No
32	Sequencer	Instrument used to produce primary data. Separate multiple entries using semicolons.	Illumina GAIIx;SOLID	Yes	No	Illumina GAIIx, SOLID, 454, ABI 3730xl

Index column indicates the order in which the columns are expected. All headers are case sensitive. The Case Sensitive column specifies which values are case sensitive. The Null column indicates which MAF columns are allowed to have null values. The Enumerated column indicates which MAF columns have specified values: an Enumerated value of "No" indicates that there are no specified values for that column; other values indicate the specific values listed allowed; a value of "Set" indicates that the MAF column values come from a specified set of known values (*e.g.* HUGO gene symbols).

MAF File Checks

The DCC Archive Validator checks the integrity of a MAF file. Validation will fail if any of the below are not true for a MAF file (Blue text indicates column header names):

1. Column header text (including case) and order must match SOP (Table 1) exactly
2. Values under column headers listed in the SOP (Table 1) as not null must have values
3. Values that are specified in Table 1 as Case Sensitive must be.
4. If column headers are listed in the SOP as having *enumerated* values (*i.e.* a "Yes" in the "Enumerated" column), then the values under those column must come from the enumerated values listed under "Enumerated".
5. If column headers are listed in the SOP as having *set* values (*i.e.* a "Set" in the "Enumerated" column), then the values under those column must come from the enumerated values of that domain (*e.g.* HUGO gene symbols).
6. All Allele-based columns must contain "nt" (not tested), - (deletion), or a string composed of the following capitalized letters: A, T, G, C.
7. If Validation_Status == "Unknown" then Tumor_Validation_Allele1, Tumor_Validation_Allele2, Match_Norm_Validation_Allele1, Match_Norm_Validation_Allele2 can be null (depending on Validation_Status).
8. If Validation_Status == Valid, then Validated_Tumor_Allele1 and Validated_Tumor_Allele2 must be populated (one of A, C, G, T, and -)
9. Verification_Status and Validation_Status should not conflict (e.g. Wildtype vs Valid).
10. Check allele values against Mutation_Status:
    a. If Mutation_Status == "Germline", then Tumor_Seq_Allele1 == Match_Norm_Seq_Allele1 and Tumor_Seq_Allele2 == Match_Norm_Seq_Allele2.
    b. If Mutation_Status == "Somatic" and Validation_Status == "Valid", then Match_Norm_Validation_Allele1 == Reference_Allele and Match_Norm_Validation_Allele2 == Reference_Allele and (Tumor_Seq_Allele1 or Tumor_Seq_Allele2) != Reference_Allele
    c. If Mutation_Status == "LOH" and Validation_Status==Unknown, then Tumor_Seq_Allele1 == Tumor_Seq_Allele2 and Match_Norm_Seq_Allele1 != Match_Norm_Seq_Allele2 and Tumor_Seq_Allele1 = (Match_Norm_Seq_Allele1 or Match_Norm_Seq_Allele2)
    d. If Mutation_Status == "LOH" and Validation_Status==Valid, then Tumor_Validation_Allele1 == Tumor_Validation_Allele2 and Match_Norm_Validation_Allele1 != Match_Norm_Validation_Allele2 and Tumor_Validation_Allele1 == (Match_Norm_Validation_Allele1 or Match_Norm_Validation_Allele2).

11. Check allele values against Validation_status:
    a. If Validation_status == “Wildtype”, then
       Tumor_Seq_Allele1=Tumor_Seq_Allele2 and
       Tumor_Seq_Allele1=Reference_Allele
12. Check that Start_position <= End_position
13. Check for the Start_position and End_position against Variant_Type:
    a. If Variant_Type == “Ins”, then
       End_position - Start_position == 1, and
       Reference_Allele == “-“, and
       (Tumor_Seq_Allele1 or Tumor_Seq_Allele2) == “-“.
    b. If Variant_Type is “Del”, then
       Reference_Allele != “-“, and
       (Tumor_Seq_Allele1 or Tumor_Seq_Allele2) == “-“.
    c. If Variant_Type != “Ins” then
       End_position - Start_position +1 == length(Reference_Allele) and
       (Tumor_Seq_Allele1 or Tumor_Seq_Allele2) ==
       length(Reference_Allele).